Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2009) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Computer Speech and Language.

Special Issue on Hybrid Machine Translation

Hybrid Arabic-French Machine Translation using Syntactic Re-ordering and Morphological Pre-processing

Emad Mohamed[1] and Fatiha Sadat[2+]

[1] *Suez Canal University, Cairo, Egypt*

[2] *University of Quebec in Montreal (UQAM)*
201 President Kennedy, Montreal, QC, Canada, H2X 3Y7

Abstract

Arabic is a highly inflected language and a morpho-syntactically complex language with many differences compared to several languages that are heavily studied. It may thus require good pre-processing as it presents significant challenges for Natural Language Processing (NLP), specifically for Machine Translation (MT). This paper aims to examine how Statistical Machine Translation (SMT) can be improved using rule-based pre-processing and language analysis. We describe a hybrid translation approach coupling an Arabic-French statistical machine translation system using the Moses decoder with additional morphological rules that reduce the morphology of the source language (Arabic) to a level that makes it closer to that of the target language (French). Moreover, we introduce additional swapping rules for a structural matching between the source language and the target language. Two structural changes involving the positions of the pronouns and verbs in both the source and target languages have been attempted. The results show an improvement in the quality of translation and a gain in terms of BLEU score after introducing a pre-processing scheme for Arabic and applying these rules based on morphological variations and verb re-ordering (VS into SV constructions) in the source language (Arabic) according to their positions in the target language (French). Furthermore, a learning curve shows the improvement in terms on BLEU score under scarce- and large-resources conditions. The proposed approach is completed without increasing the amount of training data or radically changing the algorithms that can affect the translation or training engines.

Keywords: machine translation; linguistic analysis; Arabic morphology; BLEU; Moses; Arabic-French statistical machine translation.

1. Introduction

+ Corresponding author. Tel.:+1-514-987-3000 ext. 3885.*E-mail address*: sadat.fatiha@uqam.ca.

Research in machine translation (MT) has spanned several approaches. Statistical Machine Translation (SMT) has been the approach most widely used according to a survey (Lopez, 2008). Rule-Based Machine Translation (RBMT) systems heavily depend on explicit linguistic information about source and target languages such as monolingual or bilingual dictionaries, and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively (Hutchins and Somers, 1992).

While SMT systems suffer from a lack of a real grammatical structure, often resulting in ungrammatical sentences, RBMT systems have to deal with a lack of lexical coverage. Hybrid architectures seek to combine the advantages of the individual paradigms to achieve an overall better translation (Hunsicker et al., 2012).

The idea that morphological richness may create challenges for SMT is not new. Languages with rich morphological systems present significant hurdles for SMT. Many of these languages tend to have freer word order compared to isolating languages. Moreover, words in morphologically complex languages are frequently composed of multiple morphemes in addition to a single or compound word stem, leading to a significant data sparsity problem.

Arabic is a morphologically rich and complex language, in which a white space-delimited unit, a word henceforth, may not carry only inflections but also clitics, such as pronouns, conjunctions, and prepositions. Many of these play syntactic roles like subject and object, and what looks like a word turns into a complete self-sufficient sentence. The most notable clitics are many coordinating conjunctions, the definite article, many prepositions and particles, and a class of pronouns that attach themselves either to the start or the end of words (Attia, 2008). This morphological complexity has consequences for NLP applications.

Much of the work on Statistical Machine Translation (SMT) from morphologically rich languages has shown that morphological tokenization and orthographic normalization help improve SMT quality because of the sparsity reduction they contribute (El Kholy and Habash, 2010). The complex morphology of Arabic is the main reason behind the problems of sparsity, agreement, i.e. rules that govern the inflection of words according to their relations with other words, lexical divergences and related problems. Sparsity, especially for a source language that is morphologically rich and complex, causes many errors in SMT output in a number of ways: (a) the correct inflection of a word that agrees with the rest of the sentence could be absent from the phrase table because it was not in the training data or was infrequent and therefore was filtered, (b) poor estimation of probabilities, and (c) poor language model estimation (Sultan, 2011). To illustrate, take the Arabic word ***ws>sAEdhA***, which translates into the English ***and I will help her*** and the French ***et je vais l'aider***, where one Arabic word translates into five English words and five French words. This leads to word form sparsity as forms do not occur very often (ws>sAEd**h** = and I will help **him** and ws>sAEd**hmA** = and I will help **both of them**).

In general, the more data on which a MT system is trained, the more accurately it performs. Although large amounts of parallel data exist, we think that no matter how large the data is, morphological analysis and segmentation of Arabic will be of benefit in the SMT research even if this does not mean an increase in the BLEU score.

Previous research has shown that the morphological pre-processing of morphologically rich languages, such as Arabic, can be of value especially in the case of limited volumes of training data (Goldwater and McClosky, 2005), (Sadat and Habash, 2006), (Lee, 2004), (El Ishibani et al., 2006), (Hasan et al., 2003). In the context of SMT, Habash and Sadat (Habash and Sadat, 2006) pre-processed Arabic texts using different segmentation schemes for translation into English and showed that the quality of translation is generally better than the baseline. Similar findings were reported by (El Ishibani et al., 2006) on Arabic-English SMT.

Developing a high quality Arabic-French machine translation system is not an easy task, compared to the Arabic-English or French-English pairs of languages due to the lack of freely available linguistic resources such as Arabic-French parallel corpora, the scarcity of research on this pair of languages, and the focus on Arabic-English and Chinese-English for evaluation purposes (Sasa and al.,2006). While translation from Arabic to French could also be approached by pivoting through English, this too does not seem to have attracted research attention.

While not much research on Arabic French SMT exists, the few available evaluations show a similar pattern to that of Arabic-into-English SMT. One of the first statistically-driven machine translation systems for Arabic-French was reported by Hasan et al. (Hasan et al., 2006) during the second Cesta evaluation campaign[2]. The proposed SMT system used a simple stemming algorithm based on finite-state automata to split an Arabic word into prefixes, a

[2] http://www.technolangue.net/article.php3?id_article=199

stem and suffixes. This simple segmentation method reduced the Out Of Vocabulary (OOV) words rate from 8.2% to 2.6% for the test data and thus produced a better quality of translation in terms of the BLEU score (Papineni et al., 2001). Other researches on Arabic-French SMT focused on domain adaptation to the news domain and did not consider the pre-processing of the morphologically complex language such as Arabic (Schwenk and Senellart, 2009). An improvement of 3.5 BLEU points on the test set was achieved when adapting the SMT system to the news domain by using large amounts of monolingual texts in the source language.

Morphological Pre-processing schemes have been widely adopted for handling Arabic morphology in SMT (e.g., Sadat and Habash (2006), Zollmann et al. (2006), Lee (2004)). While morphological complexity is the most observable trait of Arabic, the syntax also shows differences. In the area of machine translation, syntactic issues have not received as much attention in comparison with morphological pre-processing (Green et al. (2009), Crego and Habash (2008), Habash (2007)). Arabic verbal constructions are particularly challenging since subjects can occur in pre-verbal (SV), post-verbal (VS) or pro-dropped ("null subject") constructions (Carpuat et al., 2010).

Our analysis of the Prague Arabic Dependency Treebank (PADTB) (Hajič and Zemánek, 2004) shows that 67% of all the sentences in the PADTB are verb-first sentences. Carpuat et al. (2010) showed that post-verbal subject (VS) constructions are hard to translate because they have highly ambiguous reordering patterns when translated to English. They proposed to reorder VS construction into SV order for SMT word alignment only. This strategy significantly improved BLEU and TER scores of the SMT using Arabic-English language pair. Lately, Sadat (2013) and Sadat and Mohamed (2013) showed a study on Arabic-French machine translation introducing a pre-processing scheme for Arabic and applying some rules based on morphological variations of the source language in relation to the target language. The first proposed pre-processing scheme using morphology reduction for Arabic showed an improvement of 7.4% in terms of the BLEU score; however, the second introduced set of swapped rules dealing with pronouns showed slightly worse BLEU scores than the one with morphological processing only.

In this paper, we introduce additional swapping that re-orders the Arabic verbs in a sentence according to their positions in the French target sentence. We report on and expand our first participation in the TRAD 2012 evaluation campaign[3] that was coordinated by LNE (*Laboratoire National de métrologie et d'Essais)* and CASSIDIAN (*the Defence and Security Subsidiary of the EADS group*) and was funded by the French General Directorate for Armament (DGA). This international evaluation campaign was the first to target the Arabic-French language pair. Our main interest at this stage is to measure the effect of the pre-processing of the source language on the quality of Arabic-into-French translation, rather than how much increasing the amount of training data can improve SMT quality. However, we also report on an extensive study on the effect of Arabic-into-French SMT quality of the proposed pre-processing methods across a learning curve.

This paper is organized as follows. The morphology of Arabic language is described in Section 2. In Section 3, we discuss the proposed solutions of pre-processing Arabic through segmentation and different rules on morphological reduction of the source language. Additional rules based on syntactic re-ordering are describes in Section 3. In Section 4, we present the experiments on Arabic-French SMT with different evaluations. Section 5 concludes the present paper with a discussion and future extensions.

2. The Morphology of Arabic Language

The Arabic script is complicated in that each white-space-delimited unit may correspond to several syntactic units. An Arabic orthographic unit, a unit delimited by white space, usually carries more than one token. An example is a form like (*wsyktbwnhA*)[4] (In Eng. "*and they will write it*", depicted in Fig. 1). This grammatically complete sentence carries a conjunction ***w***, a future particle ***s***, a verbal token ***yktbwn***, and a feminine singular third person object pronoun ***hA***. The verbal token is made of a verb ***ktb***, a masculine present third person inflection ***y*** and a plural indicative inflection ***wn***. This nature entails that the type token ratio is much smaller than it is for a non-morphologically rich language like English or French, since Arabic is very rich in both vocabulary and morphological variation. It follows that any particular word will appear less often than in French or English for a

[3] http://www.trad-campaign.org/

[4] All Arabic transliterations are provided using the Buckwalter transliteration scheme (Buckwalter, 2002)

given text length and type (Hmeidi et al, 1997). In order for any linguistic, especially lexical, investigation to be reliable, one needs to perform some sort of morphological analysis capable of reducing the word to its basic form. This has implications for machine translation as it means that no matter how big the training corpus is; the Arabic side will always suffer from scarcity.

Combined with this morphological complexity, and possibly due to it, the Arabic word order is flexible. While the six order combinations of the SVO are possible, only VSO, SVO, and VOS are in actual use. This combination of morphology and syntax is what we are trying to handle in this research paper.

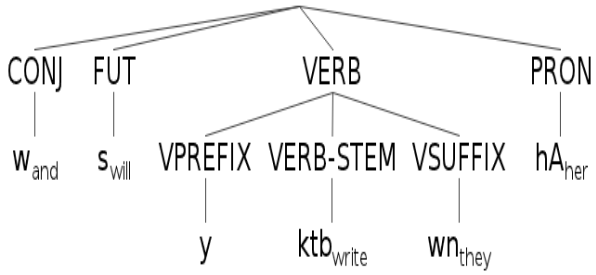

Fig. 1. The morphology of an Arabic word (In Eng. "*and they will write it*")

3. Pre-processing Arabic for SMT

With Arabic being morphologically complex and rich, lexical scarcity comes as a natural result. In such cases it helps to reduce this morphological complexity in order to obtain better alignments and decoding for SMT (Habash et al., 2010).

For Arabic pre-processing, we used the segmenter/part of speech tagger developed by Mohamed and Kübler (2010). This memory-based tagger produces fine-grained segments and part of speech tags that can with simple rules produce almost any desired level of tokenization and tag sets. Given an input sentence like (a), the system produces (b) as a segmented and annotated sentence, as described in the following example:

(a) وقد ارتبطت الاضطرابات بترحيل السلطات الفرنسية للعديد من المهاجرين غير الشرعيين

In Buckwalter transliteration: *wqd ArtbTt AlADTrAbAt btrHyl AlslTAt Alfrnsyp llEdyd mn AlmhAjryn gyr Al$rEyyn*.

In English: *The disorders have been linked to the deportation by French authorities for many illegal immigrants.*

In French: *Les troubles ont été liés à la déportation par les autorités françaises pour de nombreux immigrants clandestins.*

(b) w/**CONJ**+qd/**VERB_PART** ArtbT/**PV**+t/**PVSUFF_SUBJ:3FS**
Al/**DET**+ADTrAb/**NOUN**+At/**NSUFF_FEM_PL** b/**PREP**+trHyl/**NOUN**
Al/**DET**+slT/**NOUN**+At/**NSUFF_FEM_PL** Al/**DET**+frnsy/**ADJ**+p/**NSUFF_FEM_SG**
l/**PREP**+l/**DET**+Edyd/**NOUN** mn/**PREP** Al/**DET**+mhAjr/**NOUN**+yn/**NSUFF_MASC_PL_GEN**
gyr/**NEG_PART** Al/**DET**+$rEy/**ADJ**+yn/ **NSUFF_MASC_PL_GEN**

The Memory-based Arabic tagger has a segmentation accuracy of 98.5%, and a segmentation-tagging pipeline accuracy of 94% on the Arabic Treebank data. We have not evaluated the tagger on our current data but we expect deterioration in accuracy. We have examined a few sentences tokenized/tagged by the tagger and found out that it is more likely to miss a token boundary than to over-tokenize.

In this research paper, in addition to the **Baseline** and **Tokenized** pre-processing approaches, we propose some linguistics rules based on the variations in the output of the above example.

Note that the Baseline is generalized to the whole research we are doing. In this approach, the Arabic side undergoes minimal pre-processing in which we only separate the punctuation and remove the occasional diacritization (the short vowels). Short vowels do not normally occur in Arabic, but sometimes scattered ones are there mainly for disambiguation purposes; however since their use is not standardized and subjective, their removal usually leads to better agreement between the training and test sets.

The Tokenized approach relies on splitting the prefixes and suffixes that have a syntactic value and that usually stand as independent words in other languages. Examples of these include the possessive pronouns (*-hm*, *-h*, *-y*, *-hA*), conjunctions (*w*, *f*), and prepositions (*l-*, *k-*, *t-*). We have also chosen to split the Arabic definite article *Al* due to the perceived similarity in distribution between the Arabic and French definite articles.
The sentence above "wqd ArtbTt AlADTrAbAt btrHyl AlslTAt Alfrnsyp llEdyd mn AlmhAjryn gyr Al$rEyyn ", is thus tokenized as follows:

"**w/CONJ** qd/VERB_PART ArtbT/PV+t/PVSUFF_SUBJ:3FS Al/DET ADTrAb/NOUN+At/NSUFF_FEM_PL **b/PREP** trHyl/NOUN Al/DET slT/NOUN+At/NSUFF_FEM_PL Al/DET frnsy/ADJ+p/NSUFF_FEM_SG **l/PREP Al/DET** Edyd/NOUN mn/PREP Al/DET mhAjr/NOUN+yn/NSUFF_MASC_PL_GEN gyr/NEG_PART Al/DET $rEy/ADJ+yn/ NSUFF_MASC_PL_GEN".

Where the conjunction *w*, the prepositions *b* and *l*, and the definite article *Al* are no longer prefixes, but separate tokens. The process also normalized the definite article from *l* to *Al*, which is the more frequent form. It has to be noted that tokenization in Arabic is a morphologically involved process and not as simple as tokenization in English or French, which can usually be performed through simple rules. This is the reason that the tokenized version is not the baseline. In the Baseline approach, however, the French is tokenized. The proposed pre-processing approaches are described in the following sub-sections.

3.1. Morphological Reduction

In the morphologically reduced pre-processing approach (**MorphReduced)**, we propose to reduce the morphology of Arabic to a level that makes it closer to that of the French language and also reduce the rate of unknown words. An example of this is the dual form, which does not occur in French and has thus been transformed to the plural. The following table (Table 1) lists the most common examples of Arabic morphological reduction.

3.2. Swapping

The swapping pre-processing approach tries to introduce some structural matching between the source language (Arabic) and the target language (French). Two structural changes involving the positions of the pronouns and verbs in both the source and target languages have been attempted.

Table **1**
The most common rules for Arabic morphological reduction

Rule	Example before applying the rule	Example after applying the rule
Regular Plural Nominative → Regular Plural Accusative	mstwTn***wn***	AlmstwTn**yn**
Dual Nominative → Regular Plural Accusative	lAEb***An***	lAEb***yn***
	hn lm ylEb***n***	hm lm ylEb***wn***
Jussive Mood → Indicative Mood	hm lm ylEb***wA***	hn lm ylEb***wn***
	hmA lm ylEb**A**	hm lm ylEb***wn***

Note that each of the following swapping rules is built upon tokenization and morphological reduction.

3.2.1. **MorphSwapped**

We introduce two rules on Arabic pronouns as follows:

- While Arabic possessive pronouns follow the nouns, we have made them precede the nouns in order to match the French. For example ktAb -y (book -my) has now become (*-my book*) to match "*mon livre*" (in French).
- Furthermore, Arabic object pronouns, which follow the verb, have been made to precede it. An example is the Arabic sentence "*>nA >ryd h*" (In English. *I want it*) is now "*>nA h >ryd*" (In English. *I it want*) with the purpose of matching the French structure "*Je le veux*" (In English. *I it want*).

3.2.2. **SyntaxSwapped**

This syntactic-reordering *task* introduces some additional rules that re-order the Arabic verbs in a sentence according to their positions in the French target sentence. In French, the normal place for a verb is generally after the noun (SV construction); while Arabic sentences are classified simply according to whether or not they include a verb — regardless of where the verb is in the sentence.

- A verbal sentence (VS construction) will have a first word as a verb.
- A nominal sentence (SV) will have a first word as a noun. In this case, a verb may follow the first word that is a noun.

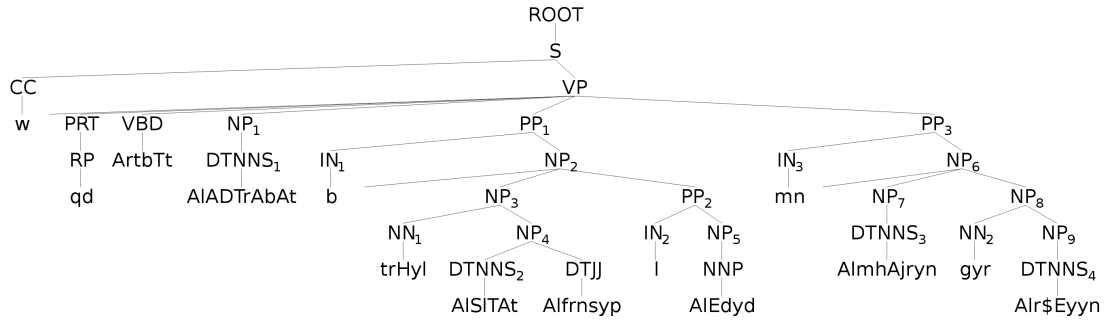

Fig. 2. An output of the Stanford parse tree based on an Arabic example

In the proposed *SyntaxSwapped* task, we re-order VS constructions into SV order for SMT. This kind of re-ordering has already shown an improvement in terms of BLEU score when applied on word alignment for SMT (Carpuat et al., 2010). We use the Stanford parser for this re-ordering, and as is with any automatic process, the parsing, and consequently the process, is not without errors. In the sentence above depicted in Fig.2 (w qd ArtbTt/VERB AlADTrAbAt/SUBJECT ...), we swap the subject and the object to create (w qd AlADTrAbAt/SUBJECT ArtbTt/VERB ...), but this is a position in which the verbal emphatic *qd* does not normally occur. There are issues like this that our re-ordering does not take care of as they require deeper syntactic and semantic analysis and could thus be a topic for future research. Unlike English, the Stanford parser does not produce functional labels like Subject and Object, and we had to determine these ourselves using the syntax tree. The Stanford output is shown for our example above in Fig. 2. We can notice that the two PP's in the sentence have been attached high to the main node, which is wrong in both cases. While this does not affect the subject extraction here, it does so in many other sentences. To extract the subject, we identify the NP that is the sister of the V directly under the main VP.

4. Experiments and Evaluations

We participated in the first participation in the TRAD 2012 evaluation campaign[5] and thus used the training, development and test data of this evaluation campaign. Our SMT system was trained on 3.5 million words of French and their parallel text in Arabic (equivalent to 108,300 sentences) in addition to 9700 parallel sentences that were extracted from the essentially comparable UN corpus of 2009. Thus, the total number of sentences is 118,000 for the training corpora. The development corpus contains 20,000 words, namely 40,000 words with the reference. The evaluation corpus contains 15,000 words with 4 references.

We followed the common practice of extracting bilingual phrases from the parallel data in three steps: first, words in the bilingual sentence pairs are aligned using GIZA++ toolkit (Och and Ney, 2003), in both directions; second, word alignment links are refined using the Grow-Diagonal-Final (GDF) heuristics. Third, bilingual phrases are extracted from the parallel data based on the refined word alignments with predefined constraints (Och and Ney, 2003). The trigram language models are implemented using the SRILM toolkit (Stolcke, 2002). We did not go beyond trigrams since that would exceed our limited computational resources, although trying higher orders of n-grams may be of research interest in its own right.

Moses[6] (Koehn et al., 2007), an open source toolkit for phrase-based SMT system, was used as a decoder. These steps of building a translation system are considered as a common practice in the state-of-the-art of phrase-based SMT systems. Our focus is placed more on the pre-processing phase than on introducing new SMT tools.

We have measured the effect of the proposed pre-processing steps on data sparseness, as measured by the percentage of Out Of Vocabulary (OOVs) unigrams in the development set. Table 2 summarizes the findings. We provide numbers in terms of tokens (the total number of words) and types (the number of unique words in the text, i.e. no-redundant words in the text). Table 4 summarizes the findings on the test set in terms of rates of OOVs. It can be noticed that tokenization has a major effect on combatting data sparseness and consequently improving the quality of translation as measured by the BLEU score. Morphological normalization, which is a layer on top of tokenization, improves things even further, and this is reflected in the difference between the Baseline BLEU score and the **MorphReduced** BLEU score, which is 8.6 absolute points. Another interesting observation is that most OOVs are related to proper nouns that were not found in our training corpora. In this case, transliteration of named entities is necessary to help improve the quality of translation.

Swapping does not seem to be uniform in its effect on SMT quality. While simple swapping rules represented by **MorphSwapped** setting lead the system output to deteriorate slightly (as compared to the **MorphReduced** Experiment), subject verb swapping represented by **SyntaxSwapped** setting showed a better result (and the best result overall) with an improvement of 8.9 absolute BLEU points above the baseline experiment.
Table 3 compares the results, in term of BLEU scores, of the five experimental settings in 3 evaluation schemes as follows:

(a) ***Standard***, which includes performing re-casing and removing white space before punctuation. For re-casing, we wrote our own machine learning re-caser using TiMBL memory-based learner toolkit with only lexical and character features. The re-caser has an accuracy of 98.5%,

(b) ***Nopunct***, in which punctuation is stripped and evaluation is performed on the lexical text only, and

(c) ***Nopunctcase*** in which, in addition to removing punctuation, all words are lowercased.

We can see from Table 3 that the Baseline experiment produces the lowest results, and that the tokenization scheme is a big leap with a 7.2 BLEU scores of improvement (25.9 vs. 33.1), which means that performing tokenization is really a necessary step for translating from Arabic, and that the morphological complexity of Arabic could be a hindrance to quality automatic translation. While tokenization leads to a considerable improvement, morphological reduction fares even better with a 7.4 BLEU score higher than the baseline. This could be due to the fact that the morphological reduction reduces the number of unknown words even further than tokenization alone.

---

5 http://www.trad-campaign.org/

6 Available for download following the link http://www.statmt.org/ moses/

MorphSwapped elements to match the target language (pronouns only), which is built upon tokenization and morphological reduction, leads to a slight deterioration of the results as it hurts the positive effect of the morphological reduction process with a loss a little over 1 BLEU score compared to tokenization alone.
Re-ordering the subject and verb to match the target language (SyntaxSwapped) is built upon tokenization and morphological reduction. It helps improve the BLEU score a little over the MorphReduced experiments in spite of the parsing problems (and hence consistency issues) mentioned above. The Stanford Arabic parser[7] does not give enough information for experimenting with full Arabic re-ordering. We plan on using a dependency parser to test whether this full re-ordering can help improve the quality of the translation even further.

We conducted more experiments with the proposed set of rules but using different training corpus sizes. The previous experiments and results of tables (2 and 3) were conducted on 10% of the whole training corpus, the UN, which means 3.5 million words of French and their parallel text in Arabic (equivalent to 108,300 sentences) in addition to 9,700 parallel sentences extracted from the essentially comparable UN corpus of 2009. Thus, the total number of sentences is 118,000 for the 10% of training corpora.

In order to measure the effect of training data size, we made use of more training data using the UN corpus (2000 to 2009), by increasing the amount of training data to 50% (590,000 parallel sentences), 75% (885,000 parallel sentences) and 100% (1,180,000 parallel sentences). The results of these experiments are summarized in Table 5. These results are on French with punctuation marks removed and all words lowercased.

Across all the experiments we have conducted, the *SyntaxSwapped* scheme performs best under all conditions. One interesting observation is that for the best performing system, that is *SyntaxSwapped*, the improvement in terms of BLEU score was very considerable under scarce-resource conditions; while under large-resource conditions (100% training data), there was less improvement in terms of BLEU score. Also, our experiments show an improvement of 6.9 BLUE points for the 100% training data versus 8.3, 9.2 and 8.9 BLEU points for the 75%, 50% and 10% training data).

While SyntaxSwapped is the clear winner in all the scenarios we have presented, in terms of the BLEU score and the reduction in the OOV rate, the Baseline setting is the fastest growing pre-processing approach. While SyntaxSwapped improved by 6.9 BLEU points from the 10% data to the 100% data, the Baseline setting improved by 8.9 points. While we do not have enough comparison data points to reach a decisive conclusion, but the difference in performance in the SyntaxSwapped experiments across the data sizes (2.4, 1.7, and 2.8, when we subtract the BLEU score at 10% from 50%, 50% from 75%, and 75% from 100%) seems to be smaller than that of the Baseline experiment (2.1, 2.6 and 4.2, respectively), which suggests that with more data added, the Baseline experiment could finally beat the SyntaxSwapped one if it had more data. This argument could further be corroborated by the figure (depicted by Fig. 3) in which the Baseline seems to have a more upward slope than the other pre-processing approaches.

Table **2**
Effect of pre-processing on the development set

Experiment	% OOV (Types)	% OOV (Tokens)	BLEU score
Baseline	10.74	4.81	17.69
Tokenized	7.99	2.00	25.84
MorphReduced	7.87	1.98	26.33
MorphSwapped	7.87	1.98	25.48
SyntaxSwapped	**7.87**	**1.98**	**26.53**

[7] http://nlp.stanford.edu/projects/arabic.shtml

Table **3**
Results in terms of BLEU score on the test set

Experiment	Baseline	Tokenized	MorphReduced	MorphSwapped	SyntaxSwapped
Standard	25.9	33.1	33.3	33.1	**33.5**
Nopunct	23.8	31.5	31.7	31.4	**31.9**
Nopunctcase	25.8	34.1	34.1	34.0	**34.7**

Table **4**
Effect of pre-processing on the test set using OOV rates

Experiment	% OOV (Types)	% OOV (Tokens)
Baseline	8.45	3.83
Tokenized	6.71	1.85
MorphReduced	6.32	1.31
MorphSwapped	6.32	1.31
SyntaxSwapped	6.32	1.31

Table **5**
Results in terms of BLEU score on the test set using more training data (learning curve)

	Training data on the *Nopunctcase*			
	10%	**50%**	**75%**	**100%**
Baseline	25.8	27.9	30.5	34.7
Tokenized	34.1	36.5	37.4	38.6
MorphReduced	34.1	36.9	38.3	40.9
MorphSwapped	34.0	36.3	37.0	38.3
SyntaxSwapped	**34.7**	**37.1**	**38.8**	**41.6**

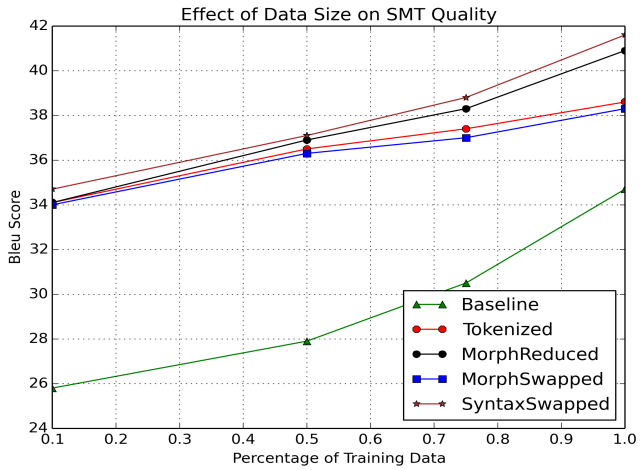

Fig. 3. Training Data Size Effect on the BLEU score

5. Conclusion

We have presented a study on hybrid machine translation for the Arabic-French pair of languages, using the training, development and test data of the TRAD 2012 evaluation campaign. We have introduced pre-processing schemes for the source language (Arabic) and some rules of language analysis related to the target language (French).

Our method that uses POS tagging and segmentation of Arabic texts showed a considerable improvement in terms of the BLEU score; however it does not achieve the best results. The introduced morphological rule that reduces the morphology of Arabic to a level that makes it closer to that of the French language showed the best results, especially when adding extra swapping rules on verb reordering (VS into SV construction), that tries to introduce some structural matching between the source language (Arabic) and the target language (French). The proposed approach is completed without increasing the amount of training data or changing radically the algorithms that can affect the translation or training engines.

A training curve showed that the effect of deeper morphological reduction and language analysis helps more small data sets than larger data sets. Moreover, this effect diminishes when adding more training data without reaching the zero, which suggests that using morphological reduction and swapping rules or re-ordering of verbs is always helpful.

Our future work is focused on the introduction of more rules for the recognition and transliteration of named entities, in order to reduce the rate of OOVs and improve more the quality of translation. We will also investigate the integration of more training data such as comparable corpora to make our MT system more competitive and reliable. A better syntactic modeling of Arabic functional heads, using a dependency parser, could be useful as it gives more control over moving heads and their dependents around, and could thus be used in fuller structural matching between French and Arabic. Finally, multi-word expressions of the source language (Arabic) to be translated to the target language (French) deserve a special attention in order to improve the quality of the whole translation.

Acknowledgements

This paper is based upon work supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors.

References

Attia, M. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. PhD Thesis. School of Languages, Linguistics and Cultures. The University of Manchester, UK (2008).

Buckwalter, T. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. Catalog: LDC2002L49 (2002).

Carpuat, M., Marton, Y. et Habash, N. Improving Arabic-to-English Statistical Machine Translation by Reordering Post-verbal Subjects for Alignment. In Proceedings of the ACL 2010 Conference Short Papers: 178–183, Uppsala, Sweden, 11-16 July (2010).

Carpuat, M., Marton, Y. et Habash, N. Reordering Matrix Post-verbal Subjects for Arabic-to-English SMT. In proceedings of the 17th Conference sur le Traitement des Langues Naturelles (TALN 2010). Montreal, Canada (2010).

Crego, J. M. and Habash, N. Using Shallow Syntax Information to Improve Word Alignment and Reordering for SMT. In Proceedings of the Third Workshop on Statistical Machine Translation : 53–61, June (2008).

Diab, M., Hacioglu, K. et Jurafsky, D. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), Boston, MA (2004).

Emad, M. et Kübler, S. Is Arabic Part of Speech Tagging Feasible Without Word Segmentation? In HLT/ACL 2010, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 705–708, Los Angeles, California, June 2010 (2010).

El Isbihani, A., Khadivi, S., Bender, O., et Ney, H. Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation. In Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-

NAACL), Workshop on Statistical Machine Translation, New York City, pages 15-22 (2006).

El Kholy, and Habash, N. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. In proceedings of TALN 2010, Montréal, July. 19–23 (2010).

Fraser, A., Weller M., Cahill, A., and Cap, F. Modeling Inflection and Word-Formation in SMT. In Proceedings of EACL (2012).

Goldwater, S. et McClosky, D. Improving Statistical MT through Morphological Analysis. In Proc. of Empirical Methods in Natural Language Processing (EMNLP), Vancouver, Canada (2005).

Green, S., Sathi, C., and Manning, C. D. NP Subject Detection in Verb-initial Arabic clauses. In Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3), (2009).

Habash, N. et Sadat, F. Arabic Preprocessing Schemes for Statistical Machine Translation. In Proceedings of NAACL 2006, New York (USA). June 5-7 (2006).

Habash, N. Syntactic pre-processing for statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*, Copenhagen (2007).

Habash, N., Rambow, O. et Ryan R. The MADA and TOKAN Manual (2010).

Hajič, J and Zemánek, P. The Prague Arabic Dependency Treebank. Development in Data and Tools (2004).

Hasan, S., El Isbihani, A. et Ney, H. Creating a Large-Scale Arabic to French Statistical Machine Translation System. In International Conference on Language resources and Evaluation (LREC), Genoa, Italy, pages 855-858 (2006).

Hmeidi I., Ghassan K. and Evens M. Design. Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. In Journal of the American Society for Information Science. 48(10): 867-881 (1997).

Hunsicker S., Yu C. and Federmann C. Machine Learning for Hybrid Machine Translation. In Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, Canada: 312–316. June 7-8, (2012).

Hutchins, W. J. and H. L. Somers. An introduction to machine translation. Academic Press, London (1992).

Koehn, P., Shen, W., Federico, M,. Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O. Zens, R., Constantin, A., Herbst, E., Moran C. et Birch, A. Moses: Open source toolkit for statistical machine translation. In Proceedings of ACL 2007 (2007).

Lee, Y. Morphological Analysis for Statistical Machine Translation. In Proc. of NAACL, Boston, MA (2004).

Lopez, A. Statistical Machine Translation. In ACM Comp. Surveys, Vol. 40, Aug (2008).

Och, F., J. et Ney, H. A Systematic Comparison of Various Statistical Alignment Models. Computational linguistics 29 (1), pages 19-51 (2003).

Manning C.D. and Schuetze H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge Mass (1999).

Papineni, K., Roukos, S., Ward, T. et Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Division, Yorktown Heights, NY (2001).

Sadat, F. et Habash, H. Arabic Preprocessing for Statistical Machine Translation: Schemes and Techniques. 2006. In Proceedings of COLING-ACL 2006, Sydney, Australia. July 17-21, (2006).

Sadat, F. Towards a Hybrid Rule-based and Statistical Arabic-French Machine Translation System. In Proceedings of Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria: 579-583. Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (Eds.) (2013).

Sadat, F and Mohamed, E. Pre-processing and Language Analysis for Arabic to French Statistical Machine Translation. In proceedings of TALN 2013, Les Sables d'Olonne. June 17-21, (2013).

Saˇsa H., El Isbihani, A, Ney H. Creating a Large-Scale Arabic to French Statistical Machine Translation System. In Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation), pp. 855-858, Genoa, Italy, May (2006).

Schwenk, H. et Senellart, J. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In MT Summit (2009).

Sultan, S. 2011. Applying Morphology to English-Arabic SMT. Master Thesis in collaboration with Google Inc. May 2011. Swiss Federal Institute of Technology, Zurich, Swiss (2011).

Stolcke, A. SRILM-An Extensible Language Modeling Toolkit. In Proceedings of the International Conference on Spoken language Processing (2002).